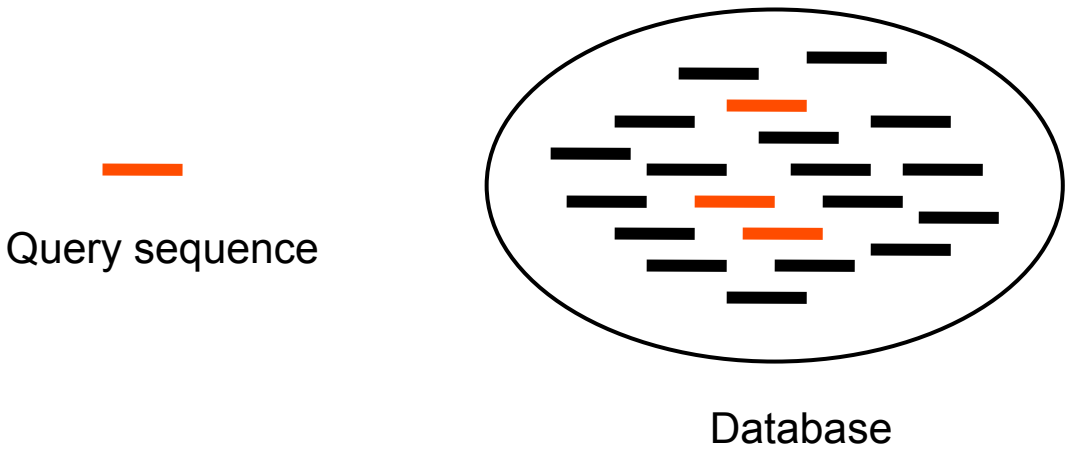
# BLAST

Ulrich Johan Kudahl

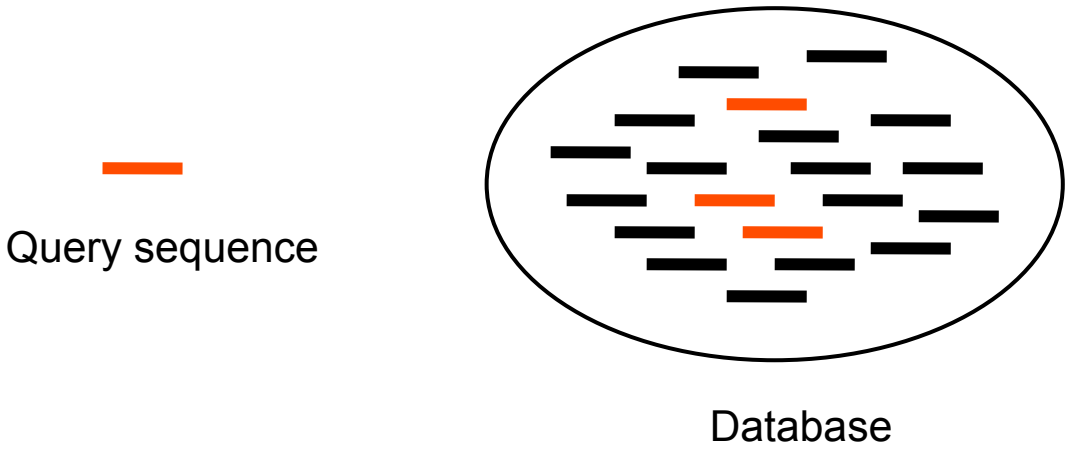
# Database searching

Using pairwise alignments to search databases for similar sequences



# Database searching

- One Pairwise alignment for each sequence
- Using Local Alignment
- Database size



# Database searching: heuristic search algorithms

BLAST (Altschul et al. 1990)

Extremely fast

Two orders of magnitude faster than Smith-Waterman

Uses statistics/likelyhood to deal with large database sizes

Very Efficient for finding rough screenings

Uses rapid word lookup methods to completely skip most of the database entries

# BLAST flavors

## BLASTN

Nucleotide query sequence

Nucleotide database

Main databases:

- NR (All sequences)
- Human

## BLASTP

Protein query sequence

Protein database

Main databases:

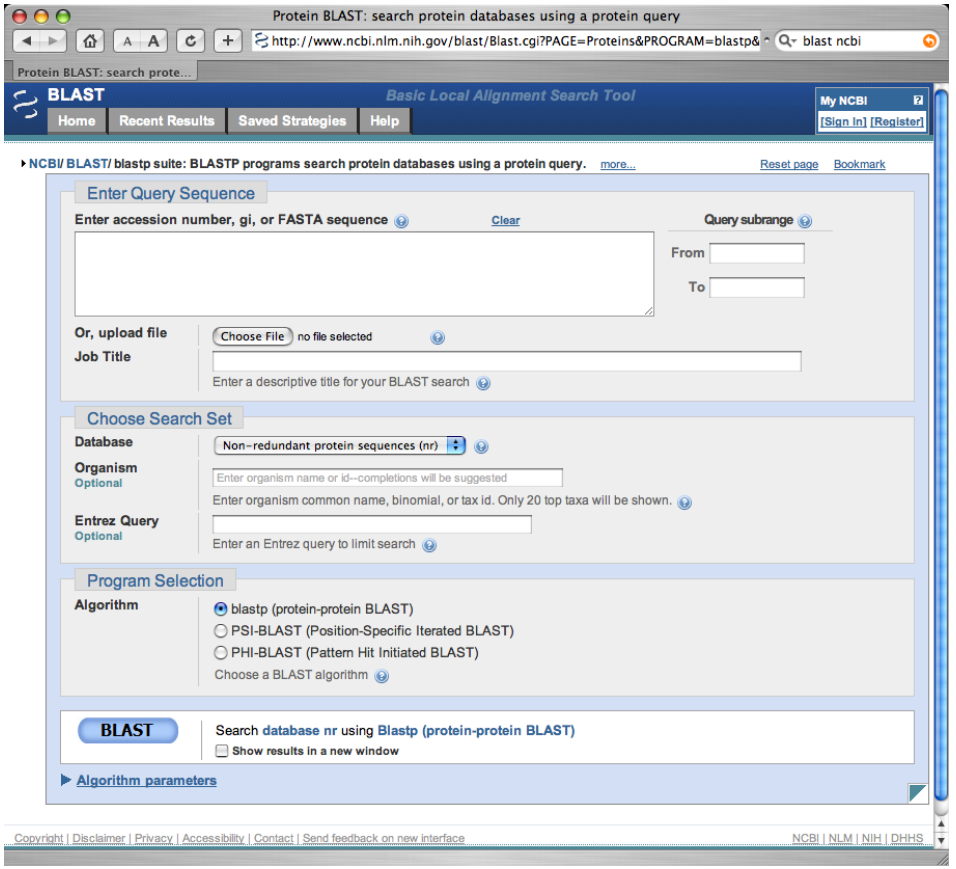
- NR (All sequences)
- Human
- Swissprot

# Searching on the web: BLAST at NCBI

Very fast computer dedicated to running BLAST searches

Many databases that are always up to date (e.g. NR and Human Genome

Nice simple web interface



# When is a database hit significant?

- Problem:
  - Even unrelated sequences can be aligned (yielding a low score)
  - How do we know if a database hit is meaningful?
  - When is an alignment score sufficiently high?

- Solution:
  - Determine the range of alignment scores you would expect to get for random reasons (i.e., when aligning unrelated sequences).
  - Compare actual scores to the distribution of random scores.
  - Is the real score much higher than you'd expect by chance?

# Distribution of random alignment scores

- Software simulation

>P29600|Savinase from Bacillus Lentus
AQSVPWGISRVQAPAAHNRGLTGSGVKVAVLDTGISTHPDLNIRGGASFVPGEPSTQDGN
GHGTHVAGTIAALNNSIGVLGVAPSAELYAVKVLGASGSGSVSSIAQGLEWAGNNGMHVA
NLSLGSPSPSATLEQAVNSATSRGVLVVAASGNSGAGSISYPARYANAMAVGATDQNNNR
ASFSQYGAGLDIVAPGVNVQSTYPGSTYASLNGTSMATPHVAGAAALVKQKNPSWSNVQI
RNHLKNTATSLGSTNLYGSGLVNAEAATR

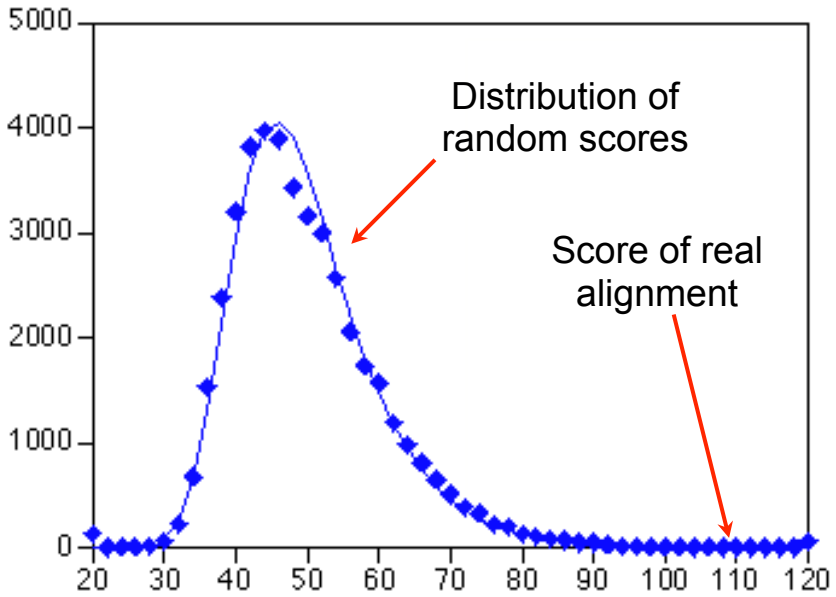
>P41363_Alkaline Protease from Bacillus Halodurans
AAKEPGAKSAQGINAGNVHKMSGGPIVISVIAIVLVNVVTAPSNYAVQNMSSASDQ
TNNTTTASHIEYLKEEYTYTDSIIVGPAIVFANSTEQNVEERQRGAKDIMQGDGVV
WGGGYAMSNGRHMFLTLRSSPSHGVLLEAANISQLGQIKSRNKTKLGGAGGSLN
ANTRIQVIANLLTNIVRYNGNFGNLNAGTSPALLTLISPINGGSSSINAPASTLAYVK
WLRGTAVSAEFRIVTVDLTAVVEHARNIGVASEVPYHPQEQHYRVAYLQGPETVNV
ASAGASDKRNFGFASSDVRGIKYVVQALTHPLYHISEHLPMDNAAGRSVDEGAAV
SAIVEYAVSLNELLGEKPANMESKE

http://www.ebi.ac.uk/Tools/psa/emboss_water/
http://www.cbs.dtu.dk/biotools/SeqShuffle-1.0/

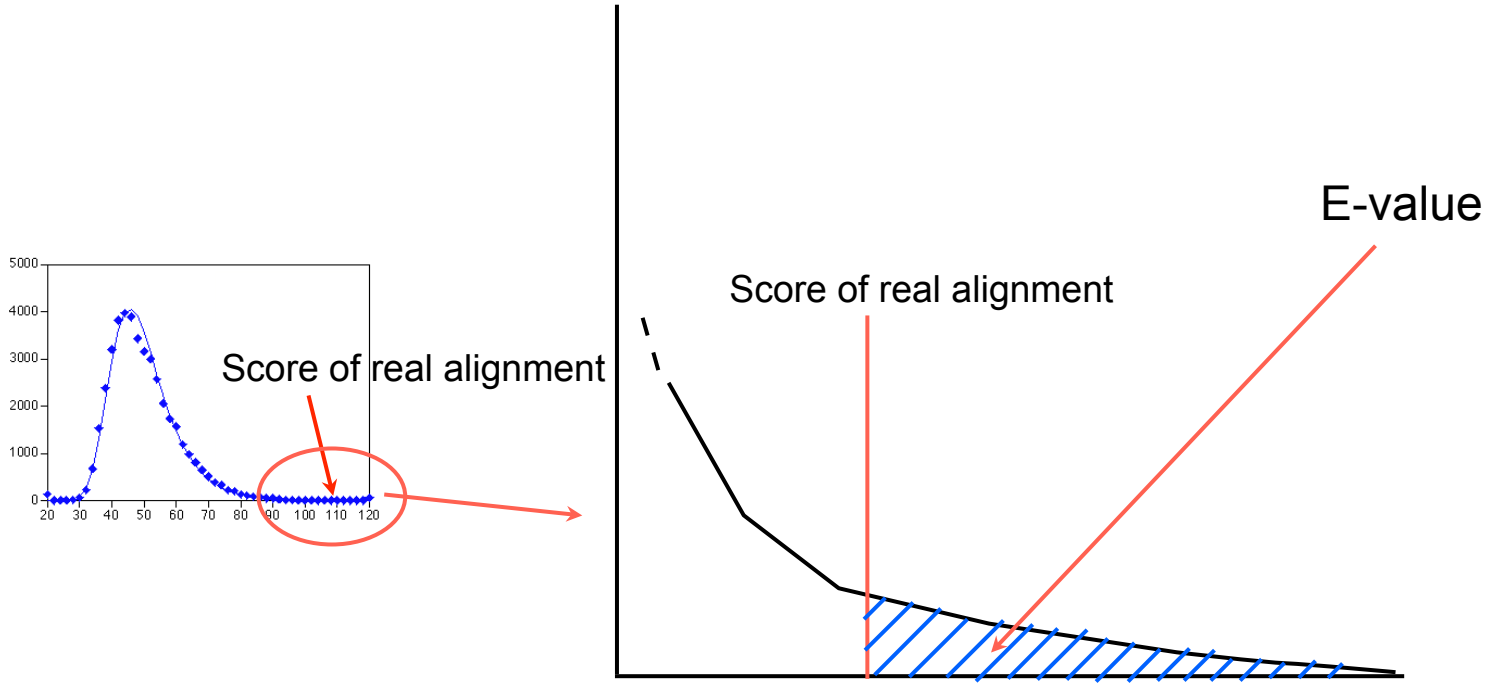
# Significance of alignment score expressed as E-value

Searching a database of unrelated sequences results in scores following an extreme value distribution

The exact shape and location of the distribution depends on the exact nature of the database and the query sequence



# Significance of alignment score expressed as E-value




E-value: the number of random hits with score ≥ real score

# BLAST heuristics

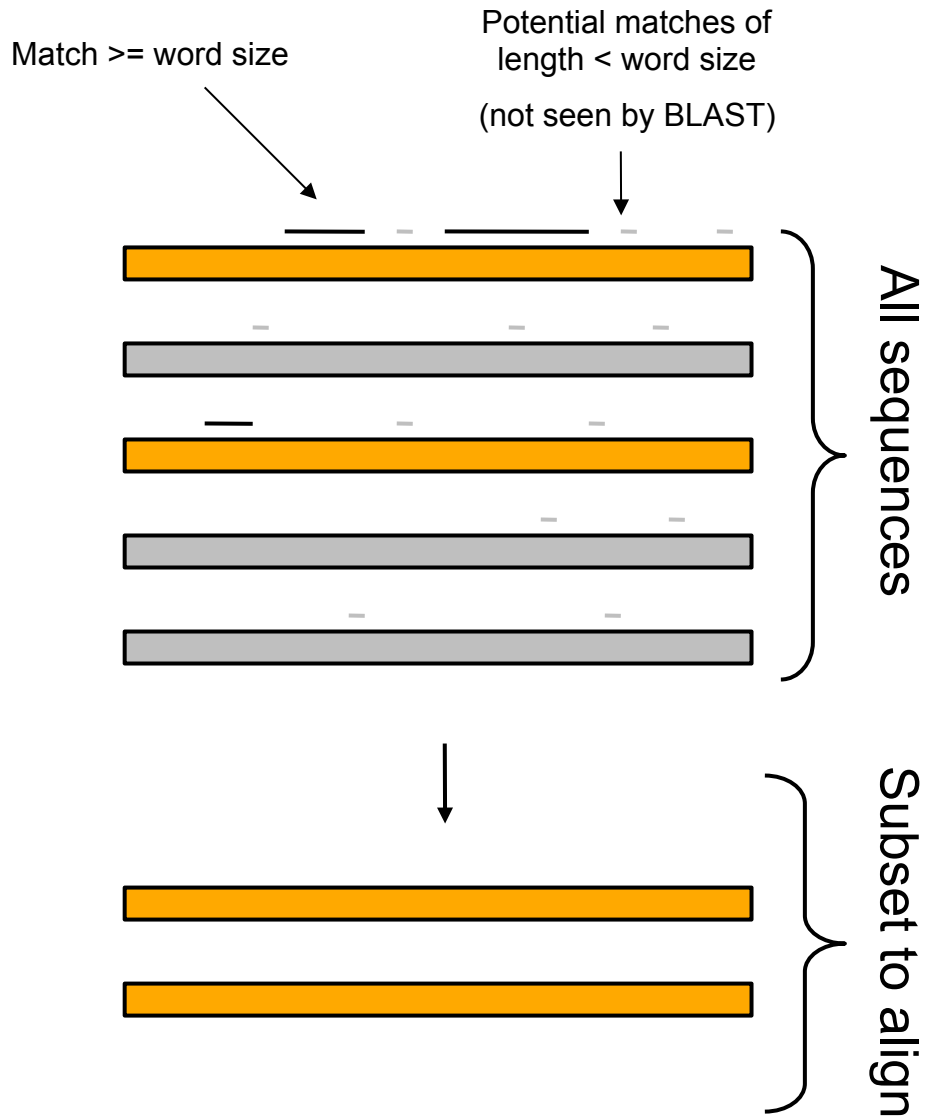
- BLAST speeds up the search >100x by pre-screening the database sequences and only performing the full Dynamic Programming on “promising” sequences.

- Promising sequences: database sequences that have sub-strings (“words”) which also occur in the query sequence (found rapidly using a so-called “suffix-tree”)

- BLASTN and BLASTP use different criteria for overlap required for a sequence to be deemed promising

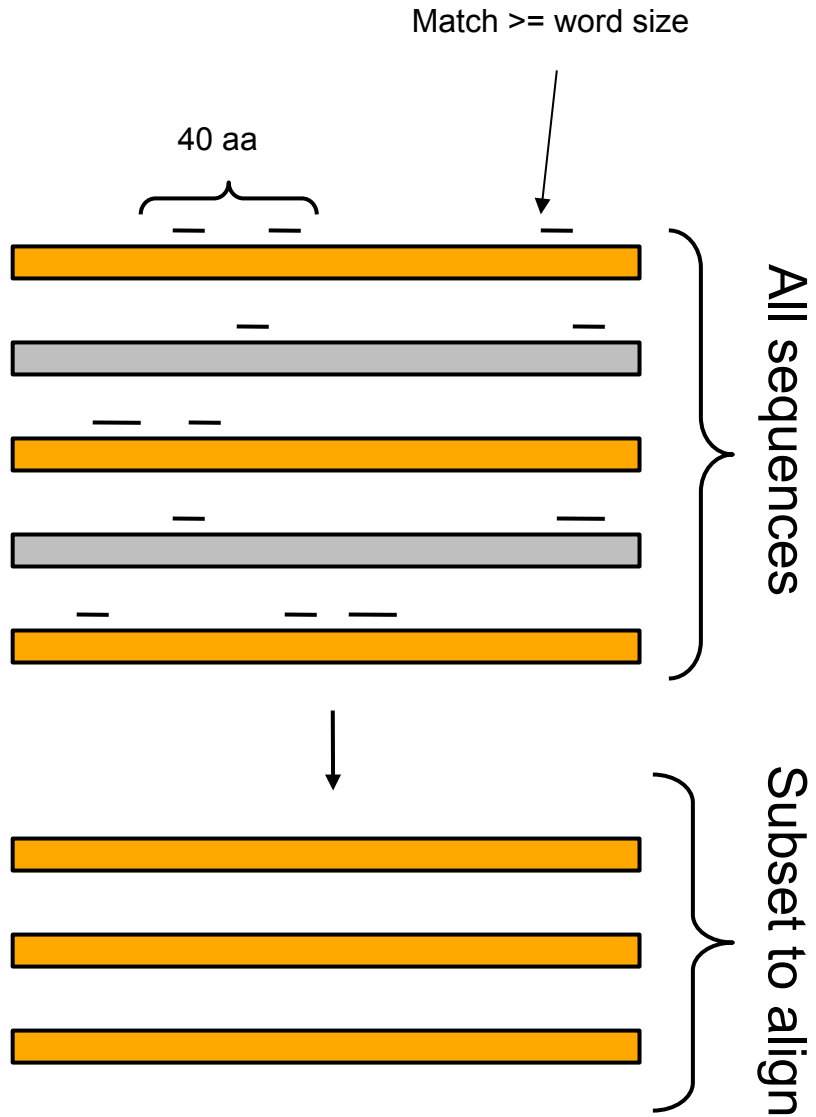
# BLASTN

- Heuristics:
  - Perfectly matching "word" of size ≥ 11 bases
- DNA alignment matrix:
  - Match: 1
  - Mismatch: -3



# BLASTP

- Heuristics:
  - 2 x “near match” within a window.
  - Default word length: 3 aa
  - Default window length: 40 aa
- Alignment matrix:
  - Default: BLOSUM 62
- Notice: These alignment matrices incorporate knowledge about protein evolution.



# BLAST

- Method for searching for DNA/protein sequences in large databases

- Blast makes a multitude of pairwise aligments and shows the best of them

- The program uses a system to evaluate the chance that the alignment is random with regard to database size and alignment length.